Assignment 03

PUBH 8878

Requirements:

- Show complete mathematical work for any derivations.
- Submit well-documented R code with clear comments and set seeds.
- Interpret results in a genetic/biological context where applicable.
- Submit the rendered PDF; do not submit the source `.qmd`.
- You may reuse and adapt your Lab 03 code, but your write-up must be self-contained.

Problem 1: EM for ABO gene frequencies — derivation, inference, and sensitivity (30 pts)

Consider phenotype counts from the ABO system under Hardy–Weinberg equilibrium (HWE): n_A, n_{AB}, n_B, n_O with total N. Let genotype frequencies be $p_A^2, 2p_Ap_O, 2p_Ap_B, p_B^2, 2p_Bp_O, p_O^2$ and $p_O = 1 - p_A - p_B$.

1) Derive the EM updates shown in lecture. Specifically, show that the E-step allocations for A and B phenotypes are

$$\tilde{n}_{AA} = n_A \frac{p_A^2}{p_A^2 + 2p_Ap_O}$$

$$\tilde{n}_{AO} = n_A \frac{2p_Ap_O}{p_A^2 + 2p_Ap_O}$$

and similarly for BB, BO, and that the M-step is

$$p_A^{(t+1)} = \frac{2\tilde{n}_{AA} + \tilde{n}_{AO} + n_{AB}}{2N}$$

$$p_B^{(t+1)} = \frac{2\tilde{n}_{BB} + \tilde{n}_{BO} + n_{AB}}{2N}$$

$$p_O^{(t+1)} = 1 - p_A^{(t+1)} - p_B^{(t+1)}$$

Hint (how to derive EM generally):

- Choose latent variables so the complete data are simple. Here, treat latent genotype counts $(n_{AA}, n_{AO}, n_{AB}, n_{BB}, n_{BO}, n_{OO})$ as missing, with only phenotype totals observed.
- Write the complete-data log-likelihood $\ell_c(p_A, p_B) = n_{AA} \log p_A^2 + n_{AO} \log(2p_A p_O) + n_{AB} \log(2p_A p_B) + n_{BB} \log p_B^2 + n_{BO} \log(2p_B p_O) + n_{OO} \log p_O^2$, using $p_O = 1 - p_A - p_B$.
- E-step: replace latent counts by their conditional expectations given the observed phenotypes under current parameters, e.g., for phenotype A, $\mathbb{E}[n_{AA} \mid A, p^{(t)}] = n_A \frac{p_A^2}{p_A^2 + 2p_A p_O}$ and $\mathbb{E}[n_{AO} \mid A, p^{(t)}] = n_A \frac{2p_A p_O}{p_A^2 + 2p_A p_O}$ (analogously for B).
- Form $Q(p \,|\, p^{(t)}) = \mathbb{E}[\ell_c(p) \mid \text{data}, p^{(t)}]$. After collecting terms, note that $Q(p) = c_A \log p_A + c_B \log p_B + c_O \log p_O + \text{const}$ where $c_A = 2\tilde{n}_{AA} + \tilde{n}_{AO} + n_{AB}$, $c_B = 2\tilde{n}_{BB} + \tilde{n}_{BO} + n_{AB}$, and $c_O = 2\tilde{n}_{OO} + \tilde{n}_{AO} + \tilde{n}_{BO}$, with $c_A + c_B + c_O = 2N$.
- M-step (intro-friendly): use the multinomial MLE fact that maximizing $c_A \log p_A + c_B \log p_B + c_O \log p_O$ under $p_A + p_B + p_O = 1$ yields normalized counts: $\hat{p}_A = c_A/(2N)$, $\hat{p}_B = c_B/(2N)$, $\hat{p}_O = c_O/(2N)$, i.e., $\hat{p}_A = (2\tilde{n}_{AA} + \tilde{n}_{AO} + n_{AB})/(2N)$, $\hat{p}_B = (2\tilde{n}_{BB} + \tilde{n}_{BO} + n_{AB})/(2N)$, and $\hat{p}_O = 1 - \hat{p}_A - \hat{p}_B$.

2) When EM breaks (haplotype non-identifiability) — a short case study. Consider two biallelic SNPs with haplotypes $\{ab, aB, Ab, AB\}$ under HWE and unphased genotypes $(g_1, g_2) \in \{0, 1, 2\}^2$.

   a. Show that if every observed genotype is $(1, 1)$, then $P\{(1, 1)\} = 2\,(p_{ab}p_{AB} + p_{aB}p_{Ab})$ and the likelihood depends only on the cross-sum $S = p_{ab}p_{AB} + p_{aB}p_{Ab}$ (a ridge; parameters not identifiable).

   b. Simulate $N = 300$ subjects with 100% $(1, 1)$. Run haplotype EM from two contrasting starts (e.g., $p^{(0)} = (0.49, 0.01, 0.01, 0.49)$ vs. $(0.01, 0.49, 0.49, 0.01)$). Report trajectories and final estimates; show equal final log-likelihoods but different limits (including boundary solutions).

   c. Add a small number of unambiguous genotypes (e.g., 10 each of $(0, 0)$ and $(2, 2)$). Re-run EM from both starts and compare. Explain how a few unambiguous observations restore identifiability and stabilize EM.

Problem 2: Two-point linkage — LOD, EM with unknown phase, and support intervals (30 pts)

Suppose one heterozygous transmitting parent (A/a and B/b) is crossed to an $aabb$ mate, yielding child haplotype counts $(n_{AB}, n_{Ab}, n_{aB}, n_{ab})$. Let $n_{\text{NR}} = n_{AB} + n_{ab}$ and $n_{\text{R}} = n_{Ab} + n_{aB}$.

1. LOD from counts (10 pts). Show that for a given recombination fraction $\theta \in (0, 0.5)$ the two-point LOD relative to independence ($\theta = 0.5$) is

$$\mathrm{LOD}(\theta) = \log_{10}\left\{ \frac{\theta^{\,n_{\mathrm{R}}}(1-\theta)^{\,n_{\mathrm{NR}}}}{0.5^{\,n_{\mathrm{R}}+n_{\mathrm{NR}}}} \right\}.$$

Derive the MLE $\hat{\theta}$ and show it equals $n_{\mathrm{R}}/(n_{\mathrm{R}} + n_{\mathrm{NR}})$ when $0 < \hat{\theta} < 0.5$.

2. 1-LOD support (8 pts). Using the direct-counting LOD, compute the 1-LOD support interval for θ by grid search over $[10^{-6}, 0.5 - 10^{-6}]$. Report the interval and briefly comment on its interpretation versus a 95% CI.

Problem 3: Two-SNP haplotype EM and LD measures (25 pts)

For two biallelic SNPs with haplotypes $\{ab, aB, Ab, AB\}$ at frequencies $\{p_{ab}, p_{aB}, p_{Ab}, p_{AB}\}$ (summing to 1), enumerate the possible haplotype pairs consistent with each unphased genotype $(g_1, g_2) \in \{0, 1, 2\}^2$.

1. (5 pts) Show that only $(g_1, g_2) = (1, 1)$ is ambiguous with two possible pairs: (ab, AB) and (aB, Ab).

Note that the E-step weights for $(1, 1)$ are proportional to $p_{ab}p_{AB}$ and $p_{aB}p_{Ab}$, and the M-step update is $p^{(t+1)} = (\text{expected hap counts})/(2N)$.

2. (10 pts) Simulate $N = 1000$ individuals from true haplotype frequencies $p^\star = (0.40, 0.10, 0.25, 0.25)$ and estimate $\hat{p}$ via EM from a uniform start. Report $\hat{p}$ and absolute errors $|\hat{p} - p^\star|$. Plot convergence of $\max_k |p_k^{(t)} - p_k^{(t-1)}|$.

3) (10pts) Compute $D = p_{11} - p_{B1}p_{B2}$ with $p_{11} = p_{AB}$, $p_{B1} = p_{Ab} + p_{AB}$, $p_{B2} = p_{aB} + p_{AB}$. Report D, D' (signed), and $r^2 = D^2/(p_{B1}(1-p_{B1})p_{B2}(1-p_{B2}))$. Comment on how LD would affect single-marker association at either SNP.

Problem 4: Single-marker association with QC and LD attenuation (15 pts)

Simulate $n = 2000$ unrelated individuals. Let a causal biallelic SNP C have MAF 0.30 and generate a quantitative trait Y with additive effect size $\beta_C = 0.50$ (per allele) and noise $\epsilon \sim \mathcal{N}(0, 1)$. Let a tag SNP T be in LD with C such that $r = \mathrm{corr}(G_C, G_T) = 0.8$ and both are in HWE.

1) Generate (G_C, G_T, Y) by first simulating haplotypes for (C, T) with a chosen LD structure that yields $r \approx 0.8$, then form genotypes and $Y = \beta_C G_C + \epsilon$. Fit simple linear models $Y \sim G_C$ and $Y \sim G_T$ and report $\hat{\beta}_C$ and $\hat{\beta}_T$, alongside their 95% confidence intervals.

2) Introduce a basic QC step: test HWE in the controls of a case–control subsample formed by thresholding Y at its 80th percentile to define cases. Compute an exact or χ^2 HWE p-value in controls for T; state whether you would flag T using a threshold of 10^{-6} and why QC is typically done in controls only.